AD-A082 995   AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TX   F/G 5/9

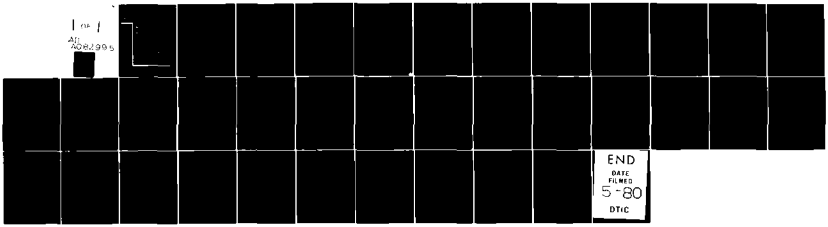
PREDICTING INVOLUNTARY SEPARATION OF ENLISTED PERSONNEL.(U)

JAN 80   W G ALBERT

UNCLASSIFIED   AFHRL-TR-79-56   NL

1 OF 1

AD A082995



AFHRL-TR-79-58

LEVEL II

AIR FORCE

HUMAN RESOURCES

LABORATORY

ADA082995

PREDICTING INVOLUNTARY SEPARATION OF ENLISTED PERSONNEL

By

Walter G. Albert

COMPUTATIONAL SCIENCES DIVISION
Brooks Air Force Base, Texas 78235

January 1980
Interim Report for Period 1 November 1977 – 30 June 1978

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This interim report was submitted by Computational Sciences Division, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas 78235, under project 6323 with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235.

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

ROBERT A. BOTTENBERG, Chief
Computational Sciences Division

RONALD W. TERRY, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

14

1. REPORT NUMBER

AFHRL-TR-79-58

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

6

4. TITLE (and Subtitle)

PREDICTING INVOLUNTARY SEPARATION OF ENLISTED PERSONNEL

9

5. TYPE OF REPORT & PERIOD COVERED

Interim rept.,

1 Nov[redacted]77 - 30 June [redacted]78

6. PERFORMING ORG. REPORT NUMBER

10

7. AUTHOR(s)

Walter G. Albert

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Computational Sciences Division

Air Force Human Resources Laboratory

Brooks Air Force Base, Texas 78235

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

62703F

16

63230511, 63230506

05

11. CONTROLLING OFFICE NAME AND ADDRESS

HQ Air Force Human Resources Laboratory (AFSC)

Brooks Air Force Base, Texas 78235

11

12. REPORT DATE

Jan[redacted]80

13. NUMBER OF PAGES

36

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

12 36

15. SECURITY CLASS. (of this report)

Unclassified

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

ABCD technique

automatic interaction detector

maximum likelihood estimation

Motivational Attrition Prediction (MAP) method

predicting involuntary separation

statistical techniques

stepwise regression

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report contains the results of a study to compare the classification accuracy of the Motivational Attrition Prediction (MAP) method to the classification accuracy of other statistical algorithms capable of predicting involuntary separation within the Air Force enlisted force. The MAP computer program, which was implemented on the UNIVAC 1108 computer system at the Air Force Human Resources Laboratory, was modified to increase its data-handling and computational capabilities and was thoroughly tested. This report includes a description of the computerized statistical algorithms, subsample selection from the first-term airman population, independent and dependent variables, model formulation and analysis, comparison of required computer resources, and related research efforts.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

404415

PREFACE

The research was completed under project 6323, Personnel Data Analyses; task 632305, Development of Analytic Methodology for Air Force Personnel Research Data.

Work Unit 63230511 was established in response to a Request for Personnel Research (RPR 77-14), entitled "Development of Improved Methods for Predicting Involuntary Separation" and initiated by the Air Force Manpower and Personnel Center.

The author wishes to express his appreciation to SrA Westley J. Richards for implementing and testing the MAP computer program on the AFHRL UNIVAC 1108 computer system and to A1C Scott I. McFarlane for increasing the program's data-handling and computational capabilities. The author also wishes to thank both of these individuals for the assistance they provided in carrying out the computer processing and data reduction phases of this study. Special acknowledgment goes to Dr. Janos B. Koplyay and Mr. C. Deene Gott for their contributions in shaping the direction of this research effort and to Mr. Larry K. Whitehead for his programming assistance.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

List of Tables (*Continued*)

PREDICTING INVOLUNTARY SEPARATION OF ENLISTED PERSONNEL

I. INTRODUCTION

In the spring of 1977, Request for Personnel Research (RPR) 77-4, Development of Enlistment Standards for the Armed Forces, was sent from the Air Force Military Personnel Center (AFMPC) [now known as the Air Force Manpower and Personnel Center] to the Air Force Human Resources Laboratory (AFHRL). The basic objectives of RPR 77-4 included the following: (a) to develop a substitute for the current Air Force enlistment standard, (b) to evaluate the Military Service Inventory (MSI) as a predictor of attrition, and (c) to test the relative efficiency of the Motivational Attrition Prediction (MAP) method in a binary classification problem, such as the prediction of retention versus attrition. Evaluation of the request by the various divisions of AFHRL eventually led to the decision to cancel RPR 77-4 and to establish two new requests, RPR 77-13, Development of Alternative Air Force Enlistment Standards, and RPR 77-14, Development of Improved Methods for Predicting Involuntary Separation. The purpose of establishing two new requests in place of the original was to facilitate the appropriate separation of research responsibility within AFHRL, i.e., the first new request dealt with objectives (a) and (b) of RPR 77-4 listed above, and the second dealt with objective (c). The Personnel Research Division (AFHRL/PE) was tasked with RPR 77-13 and the Computational Sciences Division (AFHRL/SM) with RPR 77-14.

This report describes the research carried out by AFHRL/SM in support of RPR 77-14. The basic problem concerns predicting involuntary separation (attrition) within the Air Force enlisted force. Specific objectives of this study include the following: (a) to implement the MAP computer program on the AFHRL UNIVAC 1108 computer system, (b) to compare the predictive efficiency of the MAP method with that of the AFHRL multiple linear regression technique (referred to as TRICOR), (c) to compare MAP and TRICOR with other predictive methodologies capable of handling binary criterion situations, and (d) to evaluate the various predictive methodologies using other binary criteria such as graduation/elimination from Technical Training (TT), Basic Military Training (BMT), and Undergraduate Pilot Training (UPT). This last objective (d) is not addressed here but will be the subject of a subsequent report. The results included here are restricted to predicting involuntary separation.

The next section describes the statistical methodologies compared. Three predictive methodologies associated with regression theory were considered for use in this study. These methodologies will be referred to as ordinary least squares (OLS), standardized least squares (SLS), and weighted least squares (WLS). OLS was the methodology employed in the analyses described in Section V and, hence, is discussed in the following section. SLS has been compared to MAP with regard to classification accuracy in several problem settings (Beatty, 1977). Basically, the use of standardized least squares allows the creation of a predictive model that is independent of the units of measurement since the independent variables have been normalized to zero mean and unit variance. This methodology was tested in the present study and, as expected, yielded classification accuracy results equivalent to those for OLS. An in-depth examination of the predictive efficiency of SLS will be conducted in the follow-up efforts referred to in objective (d) and discussed briefly in the last section of this report. A consideration in applying OLS to a predictive problem involving a binary criterion is that the error variances are unequal. Although the application of OLS results in unbiased estimates of the regression coefficients, the estimates are inefficient since they will not have the minimum variance property among the class of unbiased estimators. Performance of the WLS computations (Draper & Smith, 1966) results in constant error variances allowing a possible decrease in the variance associated with each estimated regression coefficient. Although WLS offers a potential improvement to OLS, its capability to accurately classify individuals as successes/failures was not examined in detail since (a) a study using a quickly assembled WLS computer programming package produced classification accuracy results similar to those for OLS, (b) some WLS analyses yielded nonsensical results, and (c) implementation of an efficient WLS computer programming package to perform analyses similar to those for OLS would not have allowed timely completion of the milestones associated with this research effort.

Following the discussion of statistical methodologies employed are sections on the airman population, description of predictive variables, selection of subsamples, model formulation and analysis, and comparison of computer resources required. Numerous tables are displayed for comparative purposes, and results and recommendations are discussed last.

II. DESCRIPTION OF STATISTICAL METHODOLOGIES

The statistical methodologies examined in this study for their ability to correctly classify individuals as successes/failures are the following: TRICOR, a computer programming package containing a stepwise regression algorithm; MAP, a computerized algorithm based on maximum likelihood estimation and utility theory; and BAYS, a computerized algorithm utilizing Bayes' formula. The stepwise regression theory of TRICOR is described in numerous publications (Dixon, 1968; Draper & Smith, 1966; Efroymson, 1960; Goldberger, 1961; Goldberger & Jochems, 1961; Pope & Webster, 1972), and the maximum likelihood estimation and utility theory of MAP is documented in AFMPC publications (Dempsey & Fast, 1976; Dempsey, Fast, & Sellman, 1977). A brief description of the important characteristics of BAYS will be presented here, and a more detailed description is available in the computer-based SMSM program documentation library at AFHRL.

Although reader familiarity with stepwise regression theory and MAP maximum likelihood estimation and utility theory is assumed, a comparison of the limitations of the computerized implementations of the two methodologies as they exist on the AFHRL UNIVAC 1108 computer system is important to researchers who want to use either of the programs. When interfaced with a compatible hit table subroutine, TRICOR has the capability to accept a data file containing information on up to 399 predictor variables and 9,999 cases per subsample. In contrast, the current version of MAP can accept a data file containing information on 20 predictor variables and the maximum number of cases allowable can be estimated by the following formula:

$$\text{NCASES} = \frac{160{,}000}{\text{NVARS} + 3}$$

where NCASES represents the number of cases and NVARS represents the number of independent variables. For example, MAP problems utilizing 5, 7, 13, or 17 independent variables allow processing of data files containing approximately 2×10^4, 1.6×10^4, 10^4, or 8×10^3 cases, respectively. An important consideration for a potential MAP user is that the program utilizes an iterative technique (Brown, 1967) to solve a system of simultaneous nonlinear equations. As will be observed in subsequent analyses, the computerized algorithm does not always converge, denying the researcher a direct comparison of the predictive accuracies of MAP versus TRICOR or BAYS.

BAYS, a computer program whose development was based on the Attribute Bayesian Classification Decision (ABCD) technique (Moonan, 1972), utilizes Bayes' formula to compute probabilities of class membership for each case, with the result that an individual is assigned to the criterion category which has the highest a posteriori probability. An important improvement to the ABCD technique was the implementation of a stepwise procedure in the model-building algorithm whereby variables can be eliminated after they have been added to the predictive scheme. Hit tables, which indicate the number of cases correctly classified, are used to select the predictor variables that most effectively discriminate among the criterion categories. At each stage of the model-building procedure, the predictive composite is formed which corresponds to the highest classification accuracy resulting from all possible additions (or deletions) of one variable to (or from) the predictive composite existing at the previous stage. As described in Section V, several random samples of the population were constructed specifically to estimate empirical probabilities.

Aside from run time constraints which will be discussed later, BAYS has the capability to accept a data file containing information on 200 independent variables having 63 categories each; however, the total number of categories for all independent variables must not exceed 2000. Since the application of the

BAYS algorithm is restricted to analysis of categorical independent and dependent variables, a categorization was performed on each independent variable with the idea of minimizing the amount of information lost in the process. This categorization requirement precluded a BAYS analysis of models containing interactive terms. At present, BAYS does not have the capability to classify individuals in an operational setting since it does not retain information on the disposition of each case. Information is retained only on the disposition of a group of cases in the form of a hit table. The performance of a proposed work effort would rectify this deficiency. In addition, the work effort proposes that BAYS be modified to utilize a variable packing factor for storing cases on a record, dynamic storage allocation, and computational shortcuts to decrease the number of data file passes.

III. FIRST-TERM AIRMAN POPULATION

The population for this study, which consisted of 11,231 airmen who entered the Air Force between April and July 1972, was selected for two major reasons: (a) the data were immediately available since the population comprised a data file prepared to support RPR 77-13, and (b) the population was characteristically similar to the one examined by Dempsey et al. (1977) which consisted of airmen who entered the Air Force between June and August 1972. In order that each case could be classified into a criterion category in a meaningful way, separation designation numbers (SDNs) were grouped and recoded in the following manner: SDNs reflecting normal separations or active duty status were recoded to a value of one and SDNs reflecting undesirable losses such as marginal productivity/inaptitude, unfitness, or unsuitability were recoded to a value of zero. This definition of the criterion categories was coordinated with AFMPC. As a result, of the 11,231 airmen in the population, 7,694 were recoded to "one" with the remaining cases recoded to "zero" (i.e., 68.5% of the cases were coded as successes, and 31.5% were coded as failures).

IV. DESCRIPTION OF PREDICTOR VARIABLES

As previously mentioned in Section III, predictor variable information was available from a data file prepared to support RPR 77-13. Information used in the creation of that file originated from a data file created for a previous work effort and the Processing and Classification of Enlistees (PACE) file at AFHRL. Complete information on the following variables was available for all airmen in the population:

1. Scores from the aptitude tests (Administrative, Mechanical, Electrical, and General) of the Armed Services Vocational Aptitude Battery (ASVAB).

2. Scores from the Armed Forces Qualification Test (AFQT).

3. Prediction of drug use admission (PDA) score (LaChar, Sparks, Larsen, and Bisbee, 1974).

4. Military Service Inventory (MSI) score (Dempsey et al., 1977).

5. Education Number of years required to reach highest level of education.

6. Dependents Coded as 0 (1) denoting number of dependents at enlistment less than or equal to 2 (greater than 2), i.e., this variable was assigned a value of 0 if the number of dependents at enlistment was less than three, and assigned a value of 1 if the number of dependents at enlistment was greater than two.

7. High school courses The following courses were coded as 1 (0) denoting completion (noncompletion):

   a. Algebra

   b. Biology

   c. Chemistry

   d. Art

e. Geometry

f. Photography

g. Physics

h. Trigonometry

i. English

j. Home Economics

8. Age – Age in years at enlistment.

Tables A1 through A13 in Appendix A present distributions, means, standard deviations, and intercorrelations of the predictor variables for the 11,231 case population. Many of the aforementioned variables were recoded (transformed) during the analysis phase of the study; however, a description of each transformation will be deferred until the next section.

V. DATA ANALYSIS

Creation and Characteristics of Subsamples

Three random samples of 1,500, 3,000, and 6,000 cases each were drawn from the population with the requirement that the three samples of each particular size contain 10%, 35%, and 50% involuntary dischargees. Each case could appear only once in each sample but could appear in more than one sample. Each of the nine samples was randomly separated into three subsamples. A schematic representation of the subsample layout is shown in Figure 1. Hereafter, the term "base rate" referred to in the figure is defined as the percentage of correct classifications that would result if all individuals in the subsample were classified into the criterion category representing normal separations or active duty status. Subsamples 3N + 1 and 3N + 2, N = 0, 1, 2, . . ., 8 were used as validation and cross-validation subsamples, respectively, in the analysis of each subsample size-base rate combination. The empirical probabilities for the BAYS computations were derived from subsamples 3N + 3, N = 0, 1, 2, . . ., 8. Although a wide range of base rates was studied in order that the subject methodologies could be compared in a variety of problem settings, attention was primarily focused on the 65% subsample base rate which closely approximates the 68.5% population base rate.

Sample #	Sample Size	Subsample #	Subsample Size	P	Q
		1	500	90	10
1	1,500	2	500	90	10
		3	500	90	10
		4	500	65	35
2	1,500	5	500	65	35
		6	500	65	35
		7	500	50	50
3	1,500	8	500	50	50
		9	500	50	50
		10	1,000	90	10
4	3,000	11	1,000	90	10
		12	1,000	90	10
		13	1,000	65	35
5	3,000	14	1,000	65	35
		15	1,000	65	35
		16	1,000	50	50
6	3,000	17	1,000	50	50
		18	1,000	50	50
		19	2,000	90	10
7	6,000	20	2,000	90	10
		21	2,000	90	10
		22	2,000	65	35
8	6,000	23	2,000	65	35
		24	2,000	65	35
		25	2,000	50	50
9	6,000	26	2,000	50	50
		27	2,000	50	50

P – base rate

Figure 1. **Subsample layout.**

Model Formulation and Analysis

The methodological comparisons began with the set of independent variables, called Variable Set I, which comprised the predictive model developed by Dempsey et al. (1977). Four additional sets of independent variables, denoted Variable Sets II–V, were examined and are listed in Table 1. Factors influencing the selection of Variable Sets II–V were the following: (a) results of analyses on Variable Set I, (b) a regression of the criterion on a large number of independent variables, (c) large increases in "turnaround" time as the number of independent variables increases associated with the BAYS computations, (d) limitations on the number of predictor variables compatible with a MAP analysis, and (e) coordination with the AFHRL focal point on RPR 77-13 concerning results of analyses supporting that research effort.

Table 1. **Sets of Independent Variables**

I	II	III	IV	V
Admin + Elec[a]	Mechanical	Mechanical	Administrative	Administrative
AFQT[b]	Electrical	Electrical	Mechanical	Mechanical
MSI[b]	General	PDA	Electrical	Electrical
EDUC[c]	MSI	EDUC[c]	General	General
Dependents	Education	Art	MSI	MSI
Age[d]	Art	Geometry	Education	EDUC[c]
	Geometry	Photography	Algebra	Algebra
	Photography	English	Biology	Biology
	English	Home Economics	Chemistry	Chemistry
			Geometry	Geometry
			Physics	Physics
			Trigonometry	Trigonometry
			English	English
			Age	Age[e]

[a]Sum (normalized) of the scores from the administrative and electrical tests of the ASVAB.

[b]Normalized score.

[c]Coded as 0 (1) denoting number of years required to reach highest level of education less than 12 (greater than or equal to 12).

[d]Coded as 0 (1) denoting age (in years) at time of enlistment less than 19 (greater than or equal to 19).

[e]Coded as 0 (1) denoting age (in years) at time of enlistment equal to 17 (greater than 17).

Tables A14 through A27 in Appendix A, which will hereafter be referred to as "hit tables," present results of the MAP, TRICOR, and BAYS methodologies applied to a validation and cross-validation subsample for each subsample size/base rate/variable set combination. An examination of the first set of results in Table A14 provides the following information. For the 500 case validation subsample from a MAP problem involving a 50% base rate, 157 individuals who were successes (i.e., assigned a criterion value of 1) were classified as successes and 173 individuals who were failures (i.e., assigned a criterion value of 0) were classified as failures. In addition, 77 individuals who were failures were classified as successes, and 93 individuals who were successes were classified as failures. Therefore, for this particular validation subsample, 330 (or 157 + 173) individuals were correctly classified and 170 (or 77 + 93) individuals were incorrectly classified. The classification accuracy for the validation subsample was 66.0% and for the cross-validation subsample was 66.4%. The remaining hit tables comprising Tables A14 through A27 can be similarly interpreted.

As can be observed from these tables, there is little difference among the methodologies in their ability to correctly classify the sampled cases into the two criterion categories. For example, the classification accuracies from applying MAP and TRICOR to the validation and cross-validation subsamples using Variable Set I differed by no more than 2% for the 18 subsample size/base rate combinations, with neither methodology exhibiting clear superiority. In fact, 15 of the 18 differences were less than 1%. For the nine validation subsamples, the classification accuracies for MAP were greater than those for TRICOR for five problems and equal for two problems, and for the nine cross-validation subsamples, the classification accuracies for MAP were greater than those for TRICOR for four problems and equal for two problems. As shown in Tables A14 and A16, the classification accuracies from applying MAP and BAYS to Variable Set I differed by no more than 3% for all subsample size/base rate combinations with a majority of the differences being less than 1%. For the nine validation subsamples, the classification accuracies for BAYS were greater than those for MAP for five problems and equal for two problems, and for the nine cross-validation subsamples, the classification accuracies for BAYS were greater than those for MAP for three problems and less than those for MAP for six problems. A similar comparison for BAYS and TRICOR can be derived from Tables A14 and A15. As before, a majority of the differences were less than

1%. For the 18 subsample size/base rate combinations, the classification accuracies for BAYS were greater than those for TRICOR for eight problems and equal for two problems. A comparison of classification accuracies among the three methodologies across all variable sets provides similar results. Regarding the performance of the algorithms as a function of base rate, sample size, or variable set, there was little difference in their abilities to correctly classify individuals as successes/failures. The application of each methodology increases classification accuracy substantially (i.e., an improvement of approximately 13% to 23%) over the base rate for the subsamples containing 50% involuntary dischargees; however, the improvement in classification accuracy decreases dramatically (i.e., an improvement of at most 11%) for the subsamples containing 35% involuntary dischargees and becomes nearly non-existent (i.e., an improvement of at most 2%) for the subsamples containing 10% involuntary dischargees. As previously mentioned, the MAP algorithm did not converge for all problems which can be witnessed by the omission of several hit tables; therefore, all comparisons between MAP and BAYS or TRICOR will refer, of course, to the problems for which the MAP algorithm converged. It should be noted that for the three 65% base rate problems utilizing Variable Set III, the TRICOR classification accuracy was better than the MAP classification accuracy in all cases; however, when contemplating the significance of this result, consideration should be given to the large number of comparisons that were made in which none of the methodologies showed clear superiority.

Using the AFHRL automatic interaction detector algorithm, AID-4 (Gott & Koplyay, 1977; Koplyay, Gott, & Elton, 1973), interactive terms were identified in an effort to gauge the improvement of classification accuracy by adding these variables to the appropriate set of predictors. As mentioned earlier, the BAYS algorithm precludes analysis of models containing interactive terms. Since the classification accuracy results of this effort were so similar to the previous results, the corresponding hit table summaries were not reproduced in this report. When interactive terms were introduced into the MAP algorithm, significant convergence problems were encountered. For example, when Variable Sets II and III were augmented with interactive terms, the MAP algorithm did not converge for each problem. Some success in achieving convergence was realized by performing MAP analyses on a subset of the augmented Variable Set III; however, a similar attempt to achieve convergence was performed on the augmented Variable Set II with little success resulting. Based on the subsample size/base rate combinations for which the MAP algorithm converged for problems with and without interactions, little predictive efficiency was gained by allowing interactive terms to be included in the model. In fact, the largest improvement observed in classification accuracy was 1.4% with most of the differences being less than 1%. Similar results were observed for TRICOR since the largest improvement in classification accuracy was 1.6% with most of the differences being less than 1%. Although the inclusion of interactive terms in these analyses did yield some increases in classification accuracy, the magnitude of the increases would not justify development of a more complicated model.

Comparison of Required Computer Resources

Although the classification accuracy results are similar for TRICOR, BAYS, and MAP, there are differences in the computer resources required to perform the computations for each methodology. All of the comparisons to be presented refer to the version of each computer program presently operational on the AFHRL UNIVAC 1108. The magnitude of the differences could vary depending on the computer system employed, and with an extensive research effort, each predictive algorithm could probably be streamlined with respect to input/output (I/O) time, central processing unit (CPU) time or mass storage required; however, the contents of this section should serve as a valuable guide for researchers who wish to estimate the computer resources required to perform each methodology on the AFHRL UNIVAC 1108 or a similar computer system without drastically modifying the computerized algorithms. If one of these methodologies is to be used repeatedly as an operational tool to solve the type of problem investigated in this report, an effort should be initiated to tailor the identified algorithm to the specific requirements of that application.

As noted earlier, an increase in the number of independent variables associated with a BAYS problem results in a dramatic increase in "turnaround" time. The total times required for BAYS processing of 6, 9,

and 14 member variable sets for 500 case subsamples were approximately 27, 42, and 65 minutes, respectively, with over 89% of those times allocated to I/O processing; moreover, an increase in the number of cases per subsample resulted in a proportionate increase in total (and I/O) processing time. The total times required for MAP processing of 6, 9, and 14 member variable sets were approximately 3% to 4%, 4% to 5%, and 7% to 10%, respectively, of the total times required to process a similar BAYS problem with the CPU times comprising approximately 77% to 92% of the total time. A direct comparison of TRICOR processing times with MAP and BAYS was not available since the TRICOR processing involved computations germane to a follow-up research effort but not required for the results herein; therefore, any estimates of TRICOR processing times should be considered overestimates. The total times required for TRICOR processing of 6, 9, and 14 member variable sets were approximately 13% to 17%, 8% to 13%, and 5% to 7%, respectively, of the total times required to process a similar BAYS problem with the CPU times comprising approximately 8% to 15% of the total time. In addition, the I/O times comprise approximately 64% to 65%, 72% to 74%, and 76% to 80% of the total times for the 500, 1000, and 2000 case subsamples, respectively.

The I/O time required for a MAP problem is small in relation to the total time required since a large amount of information is retained in mass storage, necessitating very little file handling; however, mass storage limitations restrict the size of problems that can be processed, as was reflected in an earlier discussion. The total time required to process a MAP problem surpasses the total time required to process a similar TRICOR problem for Variable Set 4 for most problems; therefore, it appears that the TRICOR algorithm becomes more efficient in relation to the MAP algorithm with respect to total time required as the number of independent variables associated with the problem increases. The CPU times required to process a BAYS or MAP problem are comparable, but the I/O times presently required to process BAYS problems limit the use of this methodology to the solution of smaller problems than could be processed by the TRICOR or MAP algorithms. Of course, for problems involving a large number of cases and predictor variables, the TRICOR algorithm presently provides a method to seek an acceptable solution within reasonable time and mass storage constraints.

VI. SUMMARY AND RECOMMENDATIONS

In order to measure the abilities of the MAP, BAYS, and TRICOR algorithms to correctly classify individuals as normal dischargees (including active duty status) or involuntary dischargees, a population of 11,231 airmen was selected that was characteristically similar to one that had served as a data base for a MAP analysis documented by Dempsey et al. (1977). The current effort is the first phase in a project to examine the capabilities of each methodology to correctly classify individuals in binary criterion situations such as graduation/elimination from various TT courses, UPT and BMT.

To examine the classification accuracies of the statistical methodologies in a variety of problem settings, subsamples were constructed so that all possible combinations of three subsample sizes (500, 1000, and 2000 cases) and base rates (50%, 65%, and 90%) could be analyzed for each set of predictor variables. Several subsets of the following variables and/or transformations of the variables were selected for development of predictive models by each methodology: (a) scores from the aptitude tests (Administrative, Mechanical, Electrical, and General) of the ASVAB, (b) AFQT score, (c) PDA score, (d) MSI score, (e) number of years required to reach highest level of education, (f) number of dependents at enlistment, (g) age in years at enlistment, and (h) high school completion of algebra, biology, chemistry, art, geometry, photography, physics, trigonometry, English, and home economics. The classification accuracies and computer resource requirements associated with the application of each statistical methodology to all subsample size/base rate/variable set combinations were compared, resulting in several general conclusions. Overall, there was very little difference among the methodologies in their ability to correctly classify individuals as successes/failures. Application of each methodology resulted in a substantial increase in classification accuracy over the base rate for the subsamples containing 50% involuntary dischargees; however, this improvement became less pronounced for the subsamples containing 35% involuntary dischargees and decreased even further for the subsamples containing 10% involuntary dischargees. The

inclusion of AID-4 identified interaction terms in the model-building process did not yield a large enough increase in classification accuracy to justify the development of a more complicated model. Convergence problems were encountered during the MAP analyses especially when some of the sets of predictive variables were augmented with interactive terms; therefore, a comparison of predictive efficiencies among all methodologies does not exist for every subsample size/base rate/variable set combination.

Although the classification accuracy results were similar, there were differences in the computer resources required to process the data for each methodology. For all analyses, the total time required to process a BAYS problem was appreciably longer than the total time required to process a similar MAP or TRICOR problem, due mainly to the large amount of I/O time associated with performing the BAYS computations. If some proposed changes to the BAYS algorithm are implemented, the I/O time required for processing a BAYS problem possibly could be reduced by one-half; however, even with this reduction, the total times associated with the BAYS problems would have greatly surpassed the times for similar MAP or TRICOR problems. Although the total time required for processing each MAP problem was appreciably less than that required for BAYS, the CPU time required for processing a MAP problem increases rather rapidly as the number of independent variables increases; consequently, it is especially important with MAP, as with the other methodologies, to employ an efficient variable selection technique. Due to mass storage limitations, an increase in the number of independent variables associated with a MAP problem causes a corresponding decrease in the maximum number of cases allowable for analysis. If the number of cases and predictor variables associated with a particular problem is large, the superior data-handling capabilities of the TRICOR regression algorithm assume added significance; in fact, TRICOR may be the only feasible method of the three to obtain a solution.

Currently, AFHRL is conducting two follow-up studies to this effort. The first of these examines the capabilities of the MAP, TRICOR, and BAYS computerized methodologies to correctly classify individuals as TT graduates/failures and the second compares the abilities of each methodology to correctly identify BMT graduates/failures. A major difference between the present and new efforts is that the test design for the TT(BMT) study requires the validation subsamples to be randomly selected from personnel who entered TT(BMT) in 1976 and the cross-validation subsamples to be randomly selected from personnel who entered TT(BMT) in 1977 rather than selecting the validation and cross-validation subsamples from the same population. Also the predictive efficiency of standardized least squares will be measured in a variety of problem settings. Since the validation and cross-validation subsamples are not necessarily homogeneous, standardized least squares predictive models which are independent of the units of measurement, may fare better than ordinary least squares predictive models. The BMT and TT research efforts should be pursued since they more closely simulate a "real world" prediction problem in that data from one time period are used to develop a model for prediction into the next time period.

REFERENCES

Beatty, T.M. *Forecasting officer losses - an examination of methods.* Randolph AFB, TX: Air Force Military Personnel Center, September 1977.

Brown, K.M. Solution of simultaneous non-linear equations. *Communications of the ACM*, 1967, **10**, 728–729.

Dempsey, J.R., & Fast, J.C. *Predicting attrition: an empirical study at the United States Air Force Academy.* Randolph AFB, TX: Air Force Military Personnel Center, March 1976.

Dempsey, J.R., Fast, J.C., & Sellman, W.S. *A method to simultaneously reduce involuntary discharges and increase the available manpower pool.* Paper presented at the Office of the Secretary of Defense (OSD)/Office of Naval Research (ONR) Attrition Conference sponsored by the Smithsonian Institute, Leesburg, Virginia, April 1977.

Dixon, W.J. *BMD: biomedical computer programs.* Berkeley, CA: University of California Press, 1968.

Draper, N.R., & Smith, H. *Applied regression analysis.* New York: Wiley, 1966.

Efroymson, M.A. Multiple regression analysis. In A. Ralston & H.S. Wilf (Eds.), *Mathematical methods for digital computers.* New York: Wiley, 1960, 191–203.

Goldberger, A.S. Stepwise least squares: residual analysis and specification error. *Journal of the American Statistical Association*, 1961, **56**, 998–1000.

Goldberger, A.S., & Jochems, D.B. Note on stepwise least squares. *Journal of the American Statistical Association*, 1961, **56**, 105–110.

Gott, C.D., & Koplyay, J.B. *Automatic interaction detector-version 4 (AID)-4 reference manual addendum I.* AFHRL-TR-77-30, AD-A042 968. Brooks AFB, TX: Computational Sciences Division, Air Force Human Resources Laboratory, July 1977.

Koplyay, J.B., Gott, C.D., & Elton, J.H. *Automatic interaction detector-version 4 (AID)-4 reference manual.* AFHRL-TR-73-17, AD-773 803. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, October 1973.

LaChar, D., Sparks, J.C., Larsen, R.M., & Bisbee, C.T. Psychometric prediction of behavioral criteria of adaptation for USAF basic trainees. *Journal of Community Psychology*, 1974, **2**(3), 268–277.

Moonan, W.J. *ABCD: A Bayesian technique for making discriminations with qualitative variables.* Paper presented at the 14th Annual Conference of the Military Testing Association, Lake Geneva, Wisconsin, September 1972.

Pope, P.T., & Webster, J.T. The use of an F-statistic in stepwise regression procedures. *Technometrics*, 1972, **14**, 327–340.

APPENDIX A: POPULATION CHARACTERISTICS AND CLASSIFICATION ACCURACY RESULTS

Table A1. **Distribution of the ASVAB Administrative Aptitude Test Scores for the First-Term Airman Population**

Score Interval (Percentile)	First-Term Airmen Falling in Score Interval	
	Number	Percent
<30	1,157	10.3
30–39	775	6.9
40–49	1,761	15.7
50–59	2,092	18.6
60–69	2,012	17.9
70–79	1,375	12.2
80–89	1,158	10.3
90–99	901	8.0

Table A2. **Distribution of the ASVAB Mechanical Aptitude Test Scores for the First-Term Airman Population**

Score Interval (Percentile)	First-Term Airmen Falling in Score Interval	
	Number	Percent
<30	793	7.1
30–39	898	8.0
40–49	1,161	10.3
50–59	2,589	23.1
60–69	2,027	18.0
70–79	1,375	12.2
80–89	1,250	11.1
90–99	1,138	10.1

Table A3. **Distribution of the ASVAB Electrical Aptitude Test Scores for the First-Term Airman Population**

Score Interval (Percentile)	First-Term Airmen Falling in Score Interval	
	Number	Percent
<30	538	4.8
30–39	616	5.5
40–49	1,728	15.4
50–59	1,969	17.5
60–69	2,059	18.3
70–79	1,103	9.8
80–89	1,820	16.2
90–99	1,398	12.4

Table A4. **Distribution of the ASVAB General Aptitude Test Scores for the First-Term Airman Population**

Score Interval (Percentile)	First-Term Airmen Falling in Score Interval	
	Number	Percent
<50	2,634	23.5
50–59	1,979	17.6
60–69	2,522	22.5
70–79	1,521	13.5
80–89	1,483	13.2
90–99	1,092	9.7

Table A5. **Distribution of the AFQT Scores for the First-Term Airman Population**

Score Interval (Percentile)	First-Term Airmen Falling in Score Interval	
	Number	Percent
<30	262	2.3
30–39	1,793	16.0
40–49	1,544	13.7
50–59	1,781	15.9
60–69	1,599	14.2
70–79	1,486	13.2
80–89	1,791	15.9
90–99	975	8.7

Table A6. **Distribution of the PDA Scores for the First-Term Airman Population**

Score Interval	First-Term Airmen Falling in Score Interval	
	Number	Percent
0–2	2,318	20.6
3–5	3,498	31.1
6–8	2,623	23.4
9–11	1,459	13.0
12–14	766	6.8
15–17	344	3.1
>17	223	2.0

Table A7. **Distribution of the MSI Scores for the First-Term Airman Population**

Score Interval	First-Term Airmen Falling in Score Interval	
	Number	Percent
0–3	3,321	29.6
4–7	4,164	37.1
8–11	2,394	21.3
12–15	936	8.3
16–19	318	2.8
>19	98	.9

Table A8. **Distribution of Education for the First-Term Airman Population**

Interval (Years)	First-Term Airmen Falling in Interval	
	Number	Percent
<12	1,542	13.7
12	8,862	78.9
13	364	3.2
14	250	2.2
>14	213	1.9

Table A9. **Distribution of Number of Dependents at Enlistment for the First-Term Airman Population**

Interval	First-Term Airmen Falling in Interval	
	Number	Percent
0–2	11,115	99.0
3–5	116	1.0

Table A10. **Distribution of Completion/Noncompletion of High School Courses for the First-Term Airman Population**

Course	Completion		Noncompletion	
	Number	Percent	Number	Percent
Algebra	8,262	73.6	2,969	26.4
Biology	8,417	74.9	2,814	25.1
Chemistry	3,511	31.3	7,720	68.7
Art	1,567	14.0	9,664	86.0
Geometry	5,597	49.8	5,634	50.2
Photography	1,653	14.7	9,578	85.3
Physics	2,045	18.2	9,186	81.8
Trigonometry	2,172	19.3	9,059	80.7
English	10,593	94.3	638	5.7
Home Economics	1,905	17.0	9,326	83.0

Table A11. **Distribution of Age at Enlistment for the First-Term Airman Population**

Interval (Years)	First-Term Airmen Falling in Score Interval	
	Number	Percent
17	1,432	12.8
18	4,126	36.7
19	2,990	26.6
20	1,452	12.9
21	609	5.4
22	331	2.9
23	137	1.2
>23	154	1.4

Table A12. **Means and Standard Deviations of the Predictive Variables for the First-Term Airman Population**

Predictive Variable	Mean	SD
Administrative	56.71	20.67
Mechanical	58.97	20.31
Electrical	62.02	20.08
General	62.03	17.95
AFQT	60.82	19.91
PDA	6.16	4.30
MSI	6.29	4.28
Education	11.93	.91
Dependents	.00	.02
Algebra	.74	.44
Biology	.75	.43
Chemistry	.31	.46
Art	.14	.35
Geometry	.50	.50
Photography	.15	.35
Physics	.18	.39
Trigonometry	.19	.40
English	.94	.23
Home Economics	.17	.38
Age	18.84	1.48

Table A13. **Intercorrelations of the Predictor Variables for the First-term Airman Population**

Predictive Variable	Admin	Mech	Elec	Gen	AFQT	PDA	MSI	Educ	DEP	ALG	BIO	Chem	Art	Geom	Photo	Phys	Trig	Eng	Homec	Age
	Intercorrelations																			
Admin	1.00	.36	.58	.71	.46	−.16	−.17	.29	.00	.33	.16	.35	−.04	.44	.04	.31	.41	.02	.01	.12
Mech		1.00	.68	.63	.63	−.09	−.09	.17	.00	.21	.07	.26	−.02	.33	.08	.26	.32	.04	−.03	.07
Elec			1.00	.74	.71	−.14	−.14	.23	−.01	.30	.11	.33	−.02	.41	.08	.36	.41	.05	−.04	.08
Gen				1.00	.63	−.14	−.15	.26	.00	.30	.15	.40	−.02	.45	.07	.34	.44	.04	−.02	.10
AFQT					1.00	−.06	−.06	.15	.00	.21	.06	.24	.00	.29	.08	.22	.28	.03	−.03	.07
PDA						1.00	.88	−.29	−.01	−.26	−.21	−.21	−.03	−.24	−.04	−.17	−.19	−.20	−.03	−.12
MSI							1.00	−.30	.00	−.26	−.21	−.19	.00	−.23	−.02	−.14	−.17	−.18	.00	−.07
Educ								1.00	−.01	.24	.20	.25	.01	.27	.03	.23	.25	.06	.02	.34
DEP									1.00	.00	−.01	.01	−.01	.00	−.01	−.01	−.01	−.02	−.01	.07
ALG										1.00	.39	.36	.03	.60	.07	.23	.29	.41	.05	.03
BIO											1.00	.26	.05	.31	.05	.12	.15	.42	.07	.04
Chem												1.00	.00	.48	.06	.43	.45	.17	.01	.11
Art													1.00	−.01	.19	.00	.00	.10	.11	.01
Geom														1.00	.05	.35	.49	.24	.03	.08
Photo															1.00	.06	.06	.10	.10	.02
Phys																1.00	.49	.12	.02	.15
Trig																	1.00	.12	.00	.12
Eng																		1.00	.11	.00
Homec																			1.00	.01
Age																				1.00

Table A14. **Hit Tables of MAP Applied to Variable Set I for Each Subsample Size – Base Rate Combination**

		Validation Actual			Cross Validation Actual		
		1		0	1		0
Subsample Size – 500	Predicted 1	157		77	163		81
Base Rate – 50%	Predicted 0	93		173	87		169
Classification Accuracy (%)			66.0			66.4	
Subsample Size – 1000	Predicted 1	332		131	327		155
Base Rate – 50%	Predicted 0	168		369	173		345
Classification Accuracy (%)			70.1			67.2	
Subsample Size – 2000	Predicted 1	650		304	642		303
Base Rate – 50%	Predicted 0	350		696	358		697
Classification Accuracy (%)			67.3			67.0	
Subsample Size – 500	Predicted 1	299		111	292		114
Base Rate – 65%	Predicted 0	26		64	33		61
Classification Accuracy (%)			72.6			70.6	
Subsample Size – 1000	Predicted 1	561		176	536		188
Base Rate – 65%	Predicted 0	89		174	114		162
Classification Accuracy (%)			73.5			69.8	
Subsample Size – 2000	Predicted 1	1109		372	1090		369
Base Rate – 65%	Predicted 0	191		328	210		331
Classification Accuracy (%)			71.8			71.0	
Subsample Size – 500	Predicted 1	447		42	443		47
Base Rate – 90%	Predicted 0	3		8	7		3
Classification Accuracy (%)			91.0			89.2	
Subsample Size – 1000	Predicted 1	894		89	898		94
Base Rate – 90%	Predicted 0	6		11	2		6
Classification Accuracy (%)			90.5			90.4	
Subsample Size – 2000	Predicted 1	1794		186	1795		189
Base Rate – 90%	Predicted 0	6		14	5		11
Classification Accuracy (%)			90.4			90.3	

Table A15. **Hit Tables of TRICOR Applied to Variable Set I for Each Subsample Size – Base Rate Combination**

		Validation Actual 1		Validation Actual 0	Cross Validation Actual 1		Cross Validation Actual 0
Subsample Size – 500	Predicted 1	192		102	181		104
Base Rate – 50%	Predicted 0	58		148	69		146
Classification Accuracy (%)			68.0			65.4	
Subsample Size – 1000	Predicted 1	326		128	315		141
Base Rate – 50%	Predicted 0	174		372	185		359
Classification Accuracy (%)			69.8			67.4	
Subsample Size – 2000	Predicted 1	576		233	575		244
Base Rate – 50%	Predicted 0	424		767	425		756
Classification Accuracy (%)			67.2			66.6	
Subsample Size – 500	Predicted 1	290		101	288		102
Base Rate – 65%	Predicted 0	35		74	37		73
Classification Accuracy (%)			72.8			72.2	
Subsample Size – 1000	Predicted 1	562		177	536		183
Base Rate – 65%	Predicted 0	88		173	114		167
Classification Accuracy (%)			73.5			70.3	
Subsample Size – 2000	Predicted 1	1078		347	1053		345
Base Rate – 65%	Predicted 0	222		353	247		355
Classification Accuracy (%)			71.6			70.4	
Subsample Size – 500	Predicted 1	448		44	444		48
Base Rate – 90%	Predicted 0	2		6	6		2
Classification Accuracy (%)			90.8			89.2	
Subsample Size – 1000	Predicted 1	894		91	898		94
Base Rate – 90%	Predicted 0	6		9	2		6
Classification Accuracy (%)			90.3			90.4	
Subsample Size – 2000	Predicted 1	1797		190	1796		191
Base Rate – 90%	Predicted 0	3		10	4		9
Classification Accuracy (%)			90.4			90.2	

Table A16. **Hit Tables of BAYS Applied to Variable Set I for Each Subsample Size – Base Rate Combination**

		Validation Actual			Cross Validation Actual		
		1		0	1		0
Subsample Size – 500	Predicted 1	189		94	182		110
Base Rate – 50%	Predicted 0	61		156	68		140
Classification Accuracy (%)			69.0			64.4	
Subsample Size – 1000	Predicted 1	346		155	333		178
Base Rate – 50%	Predicted 0	154		345	167		322
Classification Accuracy (%)			69.1			65.5	
Subsample Size – 2000	Predicted 1	736		386	703		390
Base Rate – 50%	Predicted 0	264		614	297		610
Classification Accuracy (%)			67.5			65.6	
Subsample Size – 500	Predicted 1	287		96	273		103
Base Rate – 65%	Predicted 0	38		79	52		72
Classification Accuracy (%)			73.2			69.0	
Subsample Size – 1000	Predicted 1	589		204	558		203
Base Rate – 65%	Predicted 0	61		146	92		147
Classification Accuracy (%)			73.5			70.5	
Subsample Size – 2000	Predicted 1	1181		437	1186		445
Base Rate – 65%	Predicted 0	119		263	114		255
Classification Accuracy (%)			72.2			72.0	
Subsample Size – 500	Predicted 1	450		47	448		47
Base Rate – 90%	Predicted 0	0		3	2		3
Classification Accuracy (%)			90.6			90.2	
Subsample Size – 1000	Predicted 1	893		87	891		92
Base Rate – 90%	Predicted 0	7		13	9		8
Classification Accuracy (%)			90.6			89.9	
Subsample Size – 2000	Predicted 1	1796		189	1796		196
Base Rate – 90%	Predicted 0	4		11	4		4
Classification Accuracy (%)			90.4			90.0	

Table A17. **Hit Tables of MAP Applied to Variable Set II for Each Subsample Size – Base Rate Combination**

		Validation Actual			Cross Validation Actual		
		1		0	1		0
Subsample Size – 500	Predicted 1						
Base Rate – 50%	Predicted 0			*			
Classification Accuracy (%)							
Subsample Size – 1000	Predicted 1						
Base Rate – 50%	Predicted 0			*			
Classification Accuracy (%)							
Subsample Size – 2000	Predicted 1						
Base Rate – 50%	Predicted 0			*			
Classification Accuracy (%)							
Subsample Size – 500	Predicted 1	302		109	297		132
Base Rate – 65%	Predicted 0	23		66	28		43
Classification Accuracy (%)			73.6			68.0	
Subsample Size – 1000	Predicted 1	605		217	591		228
Base Rate – 65%	Predicted 0	45		133	59		122
Classification Accuracy (%)			73.8			71.3	
Subsample Size – 2000	Predicted 1	1161		412	1162		398
Base Rate – 65%	Predicted 0	139		288	138		302
Classification Accuracy (%)			72.4			73.2	
Subsample Size – 500	Predicted 1	449		47	445		49
Base Rate – 90%	Predicted 0	1		3	5		1
Classification Accuracy (%)			90.4			89.2	
Subsample Size – 1000	Predicted 1	900		89	898		91
Base Rate – 90%	Predicted 0	0		11	2		9
Classification Accuracy (%)			91.1			90.7	
Subsample Size – 2000	Predicted 1						
Base Rate – 90%	Predicted 0			*			
Classification Accuracy (%)							

*The MAP algorithm did not converge.

Table A18. **Hit Tables of TRICOR Applied to Variable Set II for Each Subsample Size – Base Rate Combination**

		Validation Actual			Cross Validation Actual		
		1		0	1		0
Subsample Size – 500	Predicted 1	165		65	166		77
Base Rate – 50%	Predicted 0	85		185	84		173
Classification Accuracy (%)			70.0			67.8	
Subsample Size – 1000	Predicted 1	309		129	300		135
Base Rate – 50%	Predicted 0	191		371	200		365
Classification Accuracy (%)			68.0			66.5	
Subsample Size – 2000	Predicted 1	813		464	775		466
Base Rate – 50%	Predicted 0	187		536	225		534
Classification Accuracy (%)			67.4			65.4	
Subsample Size – 500	Predicted 1	295		103	289		122
Base Rate – 65%	Predicted 0	30		72	36		53
Classification Accuracy (%)			73.4			68.4	
Subsample Size – 1000	Predicted 1	603		217	594		228
Base Rate – 65%	Predicted 0	47		133	56		122
Classification Accuracy (%)			73.6			71.6	
Subsample Size – 2000	Predicted 1	1162		409	1152		396
Base Rate – 65%	Predicted 0	138		291	148		304
Classification Accuracy (%)			72.6			72.8	
Subsample Size – 500	Predicted 1	449		46	450		48
Base Rate – 90%	Predicted 0	1		4	0		2
Classification Accuracy (%)			90.6			90.4	
Subsample Size – 1000	Predicted 1	897		88	895		89
Base Rate – 90%	Predicted 0	3		12	5		11
Classification Accuracy (%)			90.9			90.6	
Subsample Size – 2000	Predicted 1	1791		181	1792		177
Base Rate – 90%	Predicted 0	9		19	8		23
Classification Accuracy (%)			90.5			90.8	

Table A19. **Hit Tables of BAYS Applied to Variable Set II for Each Subsample Size – Base Rate Combination**

		Validation Actual			Cross Validation Actual		
		1		0	1		0
Subsample Size – 500	Predicted 1	180		80	166		101
Base Rate – 50%	Predicted 0	70		170	84		149
Classification Accuracy (%)			70.0			63.0	
Subsample Size – 1000	Predicted 1	334		160	325		165
Base Rate – 50%	Predicted 0	166		340	175		335
Classification Accuracy (%)			67.4			66.0	
Subsample Size – 2000	Predicted 1	728		382	714		392
Base Rate – 50%	Predicted 0	272		618	286		608
Classification Accuracy (%)			67.3			66.1	
Subsample Size – 500	Predicted 1	277		79	256		104
Base Rate – 65%	Predicted 0	48		96	69		71
Classification Accuracy (%)			74.6			65.4	
Subsample Size – 1000	Predicted 1	593		204	590		217
Base Rate – 65%	Predicted 0	57		146	60		133
Classification Accuracy (%)			73.9			72.3	
Subsample Size – 2000	Predicted 1	1180		417	1158		425
Base Rate – 65%	Predicted 0	120		283	142		275
Classification Accuracy (%)			73.2			71.6	
Subsample Size – 500	Predicted 1	449		44	448		47
Base Rate – 90%	Predicted 0	1		6	2		3
Classification Accuracy (%)			91.0			90.2	
Subsample Size – 1000	Predicted 1	892		81	890		83
Base Rate – 90%	Predicted 0	8		19	10		17
Classification Accuracy (%)			91.1			90.7	
Subsample Size – 2000	Predicted 1	1796		187	1797		186
Base Rate – 90%	Predicted 0	4		13	3		14
Classification Accuracy (%)			90.4			90.6	

Table A20. **Hit Tables of MAP Applied to Variable Set III for Each Subsample Size – Base Rate Combination**

		Validation Actual 1		Validation Actual 0	Cross Validation Actual 1		Cross Validation Actual 0
Subsample Size – 500	Predicted 1						
Base Rate – 50%	Predicted 0			*			
Classification Accuracy (%)							
Subsample Size – 1000	Predicted 1						
Base Rate – 50%	Predicted 0			*			
Classification Accuracy (%)							
Subsample Size – 2000	Predicted 1						
Base Rate – 50%	Predicted 0			*			
Classification Accuracy (%)							
Subsample Size – 500	Predicted 1	294		97	284		108
Base Rate – 65%	Predicted 0	31		78	41		67
Classification Accuracy (%)			74.4			70.2	
Subsample Size – 1000	Predicted 1	600		220	605		231
Base Rate – 65%	Predicted 0	50		130	45		119
Classification Accuracy (%)			73.0			72.4	
Subsample Size – 2000	Predicted 1	1273		667	1276		668
Base Rate – 65%	Predicted 0	27		33	24		32
Classification Accuracy (%)			65.3			65.4	
Subsample Size – 500	Predicted 1	450		43	448		44
Base Rate – 90%	Predicted 0	0		7	2		6
Classification Accuracy (%)			91.4			90.8	
Subsample Size – 1000	Predicted 1	900		86	895		84
Base Rate – 90%	Predicted 0	0		14	5		16
Classification Accuracy (%)			91.4			91.1	
Subsample Size – 2000	Predicted 1	1798		182	1796		183
Base Rate – 90%	Predicted 0	2		18	4		17
Classification Accuracy (%)			90.8			90.6	

*The MAP algorithm did not converge.

Table A21. **Hit Tables of TRICOR Applied to Variable Set III for Each Subsample Size – Base Rate Combination**

		Validation Actual			Cross Validation Actual		
		1		0	1		0
Subsample Size – 500	Predicted 1	191		82	185		92
Base Rate – 50%	Predicted 0	59		168	65		158
Classification Accuracy (%)			71.8			68.6	
Subsample Size – 1000	Predicted 1	317		129	318		142
Base Rate – 50%	Predicted 0	183		371	182		358
Classification Accuracy (%)			68.8			67.6	
Subsample Size – 2000	Predicted 1	731		350	685		349
Base Rate – 50%	Predicted 0	269		650	315		651
Classification Accuracy (%)			69.0			66.8	
Subsample Size – 500	Predicted 1	303		99	292		109
Base Rate – 65%	Predicted 0	22		76	33		66
Classification Accuracy (%)			75.8			71.6	
Subsample Size – 1000	Predicted 1	578		195	581		203
Base Rate – 65%	Predicted 0	72		155	69		147
Classification Accuracy (%)			73.3			72.8	
Subsample Size – 2000	Predicted 1	1097		331	1073		334
Base Rate – 65%	Predicted 0	203		369	227		366
Classification Accuracy (%)			73.3			72.0	
Subsample Size – 500	Predicted 1	446		39	446		41
Base Rate – 90%	Predicted 0	4		11	4		9
Classification Accuracy (%)			91.4			91.0	
Subsample Size – 1000	Predicted 1	900		87	897		86
Base Rate – 90%	Predicted 0	0		13	3		14
Classification Accuracy (%)			91.3			91.1	
Subsample Size – 2000	Predicted 1	1793		181	1795		177
Base Rate – 90%	Predicted 0	7		19	5		23
Classification Accuracy (%)			90.6			90.9	

Table A22. **Hit Tables of BAYS Applied to Variable Set III for Each Subsample Size – Base Rate Combination**

		Validation Actual 1		Validation Actual 0	Cross Validation Actual 1		Cross Validation Actual 0
Subsample Size - 500	Predicted 1	198		83	181		98
Base Rate - 50%	Predicted 0	52		167	69		152
Classification Accuracy (%)			73.0			66.6	
Subsample Size – 1000	Predicted 1	374		176	367		190
Base Rate - 50%	Predicted 0	126		324	133		310
Classification Accuracy (%)			69.8			67.7	
Subsample Size – 2000	Predicted 1	725		347	681		353
Base Rate - 50%	Predicted 0	275		653	319		647
Classification Accuracy (%)			68.9			66.4	
Subsample Size - 500	Predicted 1	291		102	291		115
Base Rate – 65%	Predicted 0	34		73	34		60
Classification Accuracy (%)			72.8			70.2	
Subsample Size - 1000	Predicted 1	580		189	568		199
Base Rate – 65%	Predicted 0	70		161	82		151
Classification Accuracy (%)			74.1			71.9	
Subsample Size - 2000	Predicted 1	1166		397	1156		402
Base Rate - 65%	Predicted 0	134		303	144		298
Classification Accuracy (%)			73.4			72.7	
Subsample Size - 500	Predicted 1	444		34	439		40
Base Rate - 90%	Predicted 0	6		16	11		10
Classification Accuracy (%)			92.0			89.8	
Subsample Size – 1000	Predicted 1	898		86	894		86
Base Rate - 90%	Predicted 0	2		14	6		14
Classification Accuracy (%)			91.2			90.8	
Subsample Size - 2000	Predicted 1	1791		178	1795		176
Base Rate - 90%	Predicted 0	9		22	5		24
Classification Accuracy (%)			90.6			91.0	

Table A23. **Hit Tables of MAP Applied to Variable Set IV for Each Subsample Size – Base Rate Combination**

		Validation Actual 1		Validation Actual 0	Cross Validation Actual 1		Cross Validation Actual 0
Subsample Size – 500	Predicted 1	183		87	177		85
Base Rate – 50%	Predicted 0	67		163	73		165
Classification Accuracy (%)			69.2			68.4	
Subsample Size – 1000	Predicted 1						
Base Rate – 50%	Predicted 0			*			
Classification Accuracy (%)							
Subsample Size – 2000	Predicted 1	705		376	711		370
Base Rate – 50%	Predicted 0	295		624	289		630
Classification Accuracy (%)			66.4			67.0	
Subsample Size – 500	Predicted 1	274		85	271		105
Base Rate – 65%	Predicted 0	51		90	54		70
Classification Accuracy (%)			72.8			68.2	
Subsample Size – 1000	Predicted 1	608		220	602		241
Base Rate – 65%	Predicted 0	42		130	48		109
Classification Accuracy (%)			73.8			71.1	
Subsample Size – 2000	Predicted 1	1186		447	1190		443
Base Rate – 65%	Predicted 0	114		253	110		257
Classification Accuracy (%)			72.0			72.4	
Subsample Size – 500	Predicted 1	449		46	449		47
Base Rate – 90%	Predicted 0	1		4	1		3
Classification Accuracy (%)			90.6			90.4	
Subsample Size – 1000	Predicted 1						
Base Rate – 90%	Predicted 0			*			
Classification Accuracy (%)							
Subsample Size – 2000	Predicted 1	1789		175	1786		168
Base Rate – 90%	Predicted 0	11		25	14		32
Classification Accuracy (%)			90.7			90.9	

*The MAP algorithm did not converge.

Table A24. **Hit Tables of TRICOR Applied to Variable Set IV for Each Subsample Size – Base Rate Combination**

		Validation Actual			Cross Validation Actual		
		1		0	1		0
Subsample Size – 500	Predicted 1	171		76	152		76
Base Rate – 50%	Predicted 0	79		174	98		174
Classification Accuracy (%)			69.0			65.2	
Subsample Size – 1000	Predicted 1	343		164	339		170
Base Rate – 50%	Predicted 0	157		336	161		330
Classification Accuracy (%)			67.9			66.9	
Subsample Size – 2000	Predicted 1	615		274	609		287
Base Rate – 50%	Predicted 0	385		726	391		713
Classification Accuracy (%)			67.0			66.1	
Subsample Size – 500	Predicted 1	262		77	254		93
Base Rate – 65%	Predicted 0	63		98	71		82
Classification Accuracy (%)			72.0			67.2	
Subsample Size – 1000	Predicted 1	609		224	605		242
Base Rate – 65%	Predicted 0	41		126	45		108
Classification Accuracy (%)			73.5			71.3	
Subsample Size – 2000	Predicted 1	1171		424	1165		418
Base Rate – 65%	Predicted 0	129		276	135		282
Classification Accuracy (%)			72.4			72.4	
Subsample Size – 500	Predicted 1	449		46	448		45
Base Rate – 90%	Predicted 0	1		4	2		5
Classification Accuracy (%)			90.6			90.6	
Subsample Size – 1000	Predicted 1	899		91	897		91
Base Rate – 90%	Predicted 0	1		9	3		9
Classification Accuracy (%)			90.8			90.6	
Subsample Size – 2000	Predicted 1	1796		183	1794		178
Base Rate – 90%	Predicted 0	4		17	6		22
Classification Accuracy (%)			90.6			90.8	

Table A25. **Hit Tables of BAYS Applied to Variable Set IV (or V*) for Each Subsample Size – Base Rate Combination**

		Validation Actual 1		Validation Actual 0	Cross Validation Actual 1		Cross Validation Actual 0
Subsample Size – 500	Predicted 1	181		79	165		88
Base Rate – 50%	Predicted 0	69		171	85		162
Classification Accuracy (%)			70.4			65.4	
Subsample Size – 1000	Predicted 1	345		165	353		188
Base Rate – 50%	Predicted 0	155		335	147		312
Classification Accuracy (%)			68.0			66.5	
Subsample Size – 2000	Predicted 1	694		349	674		340
Base Rate – 50%	Predicted 0	306		651	326		660
Classification Accuracy (%)			67.2			66.7	
Subsample Size – 500	Predicted 1	281		89	280		113
Base Rate – 65%	Predicted 0	44		86	45		62
Classification Accuracy (%)			73.4			68.4	
Subsample Size – 1000	Predicted 1	585		199	576		204
Base Rate – 65%	Predicted 0	65		151	74		146
Classification Accuracy (%)			73.6			72.2	
Subsample Size – 2000	Predicted 1	1142		374	1111		390
Base Rate – 65%	Predicted 0	158		326	189		310
Classification Accuracy (%)			73.4			71.0	
Subsample Size – 500	Predicted 1	449		43	447		44
Base Rate – 90%	Predicted 0	1		7	3		6
Classification Accuracy (%)			91.2			90.6	
Subsample Size – 1000	Predicted 1	895		83	889		85
Base Rate – 90%	Predicted 0	5		17	11		15
Classification Accuracy (%)			91.2			90.4	
Subsample Size – 2000	Predicted 1	1792		181	1794		177
Base Rate – 90%	Predicted 0	8		19	6		23
Classification Accuracy (%)			90.6			90.8	

*Categorizing the predictive variables in Variable Sets IV and V resulted in identical sets of variables for BAYS analyses.

Table A26. **Hit Tables of MAP Applied to Variable Set V for Each Subsample Size – Base Rate Combination**

		Validation Actual			Cross Validation Actual		
		1		0	1		0
Subsample Size – 500	Predicted 1	190		95	177		88
Base Rate – 50%	Predicted 0	60		155	73		162
Classification Accuracy (%)			69.0			67.8	
Subsample Size – 1000	Predicted 1						
Base Rate – 50%	Predicted 0			*			
Classification Accuracy (%)							
Subsample Size – 2000	Predicted 1	784		438	768		449
Base Rate – 50%	Predicted 0	216		562	232		551
Classification Accuracy (%)			67.3			66.0	
Subsample Size – 500	Predicted 1	273		80	269		98
Base Rate – 65%	Predicted 0	52		95	56		77
Classification Accuracy (%)			73.6			69.2	
Subsample Size – 1000	Predicted 1	606		224	603		244
Base Rate – 65%	Predicted 0	44		126	47		106
Classification Accuracy (%)			73.2			70.9	
Subsample Size – 2000	Predicted 1	1178		424	1157		431
Base Rate – 65%	Predicted 0	122		276	143		269
Classification Accuracy (%)			72.7			71.3	
Subsample Size – 500	Predicted 1	449		48	450		48
Base Rate – 90%	Predicted 0	1		2	0		2
Classification Accuracy (%)			90.2			90.4	
Subsample Size – 1000	Predicted 1	899		89	897		88
Base Rate – 90%	Predicted 0	1		11	3		12
Classification Accuracy (%)			91.0			90.9	
Subsample Size – 2000	Predicted 1	1796		186	1795		182
Base Rate – 90%	Predicted 0	4		14	5		18
Classification Accuracy (%)			90.5			90.6	

*The MAP algorithm did not converge.

Table A27. **Hit Tables of TRICOR Applied to Variable Set V for Each Subsample Size – Base Rate Combination**

		Validation Actual 1		Validation Actual 0	Cross Validation Actual 1		Cross Validation Actual 0
Subsample Size – 500	Predicted 1	171		70	156		69
Base Rate – 50%	Predicted 0	79		180	94		181
Classification Accuracy (%)			70.2			67.4	
Subsample Size – 1000	Predicted 1	334		153	337		161
Base Rate – 50%	Predicted 0	166		347	163		339
Classification Accuracy (%)			68.1			67.6	
Subsample Size – 2000	Predicted 1	737		397	730		391
Base Rate – 50%	Predicted 0	263		603	270		609
Classification Accuracy (%)			67.0			67.0	
Subsample Size – 500	Predicted 1	270		80	267		104
Base Rate – 65%	Predicted 0	55		95	58		71
Classification Accuracy (%)			73.0			67.6	
Subsample Size – 1000	Predicted 1	603		220	602		241
Base Rate – 65%	Predicted 0	47		130	48		109
Classification Accuracy (%)			73.3			71.1	
Subsample Size – 2000	Predicted 1	1143		395	1135		399
Base Rate – 65%	Predicted 0	157		305	165		301
Classification Accuracy (%)			72.4			71.8	
Subsample Size – 500	Predicted 1	448		45	448		45
Base Rate – 90%	Predicted 0	2		5	2		5
Classification Accuracy (%)			90.6			90.6	
Subsample Size – 1000	Predicted 1	899		90	897		89
Base Rate – 90%	Predicted 0	1		10	3		11
Classification Accuracy (%)			90.9			90.8	
Subsample Size – 2000	Predicted 1	1785		176	1782		171
Base Rate – 90%	Predicted 0	15		24	18		29
Classification Accuracy (%)			90.4			90.6	